Published in final edited form as:
Psychiatr Ann. 2008 December 1; 38(12): 784–792.

Missing Data in Longitudinal Clinical Trials Part A: Design and Conceptual Issues

Philip W. Lavori, PhD [Professor, Chair],
Biostatistics and Department of Health Research and Policy, Stanford University School of Medicine.

C. Hendricks Brown, PhD,
Prevention Science and Methodology Group, Departments of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa.

Naihua Duan, PhD [Professor, Director],
Biostatistics in Psychiatry, Departments of Biostatistics and Psychiatry, Columbia University, New York and Division of Biostatistics, N.Y. Psychiatric Institute.

Robert D. Gibbons, PhD [Professor, Director], and
Biostatistics and Psychiatry and Center for Health Statistics, University of Illinois at Chicago.

Joel Greenhouse, PhD
Department of Statistics, Carnegie Mellon University, Pittsburgh.

There are numerous types of missing data that can occur in clinical trials. Some types of missing data cannot be prevented and are beyond the research team's control. For example, a patient may relocate and be unavailable for an assessment. Other types occur because the research team actually designs the study to generate incomplete data. For example, it may not be cost-effective to obtain a full diagnostic assessment on each subject. Instead, an inexpensive screening tool is used to assess everyone. All screen positives, plus a random sample of screen negatives, are then given the full diagnostic. The incomplete diagnostic data on the majority of those screened negative are taken into account in assessing treatment effects.[1-3] These types of data are missing by design, and there are statistical procedures to handle such planned missing data if designed appropriately. The third type occurs because of a faulty design plan by the research team. In this case, data are lost because a less than adequate protocol is followed. For example, if no screen negatives are given the full diagnostic, it will not be possible to ascertain the false negatives.

As another example, we will see that serious missing data problems result from a trial design protocol that stops any further longitudinal assessment once a patient stops adhering to his or her assigned treatment. This article discusses how to avoid such faulty designs and how to design studies that can improve the design efficiency by using intentionally incomplete data. It also discusses how some analytic methods, particularly last observation carried forward (LOCF) compound, rather than resolve missing data issues.

Intentionality is a hallmark of a quality randomized trial design. A useful analogy is in the construction of a building. All buildings, be they single family homes, chemical processing plants, or sports arenas, have the general property of keeping the outside out and the inside in.

Address correspondence to: Philip Lavori, HRP/Redwood Building, 259 Campus Drive, Stanford, CA 94305-5405; fax 650-725-6951; or e-mail lavori@stanford.edu..

Dr. Lavori, Dr. Brown, Dr. Duan, Dr. Gibbons, and Dr. Greenhouse have disclosed no relevant financial relationships.

But an architect must have the building's intended use in mind, so that its entire structure conforms to those needs. The architects designing a clinical trial have the same need to go beyond describing the obvious intention to “compare treatments on an outcome, and spell out explicitly the primary research goal intended for the study. Is the primary goal to compare what would happen if everyone took drug A with what would happen if everyone took drug B? Or is the primary goal to compare the following two procedures: “offer everyone new drug A and if the patient refuses to take the drug or is nonresponsive on the drug, then give the standard drug, B” versus “offer only the standard drug, B”? Once these goals are established (or at least ordered in priority), then design choices follow. Such

issues as what is randomized, what is fixed, what is allowed to vary outside of control, what is the analysis plan, can fall naturally once the research questions and goals are carefully articulated.

The validity of the study design is defined by how closely it is likely to fulfill the primary goal. A study design, including the sampling strategy, assignment procedure, follow-up procedures, and analytical method, may be prone to systematically miss its goal, on average underestimating (or overestimating), the true but unknown quantity it is intended to evaluate. This is the design bias, alluding to the corresponding statistical concept of a biased estimator. For example, a trial that excluded cases that experienced a side effect would be expected to produce an overly optimistic estimate of an active drug's effect compared with placebo, if the side effect is negatively associated with the primary outcome of interest. Certain side effects may be positively related to outcome. For example, increased bioavailability of the drug may increase the beneficial effects of the drug but may also increase side effects. In this case, exclusion of cases that experienced a side effect would produce an overly pessimistic estimate of an active drug's effect compared to placebo. A study design may also tend to produce highly variable results, translating to a lack of confidence in their statistical stability, or a lack of precision. Such a situation occurs when the study involves too few subjects. The ultimate success of the completed trial depends on how closely it conforms to the intended design and how well the intended design addresses the primary goal, combining bias, and variability.

Bias and precision of the design depend on the realities of study conduct in the real world. One of those realities is the phenomenon of “missing data,” the set of observations planned but not obtained. A design may be unbiased and precise in the fictional setting where all observations are complete but may fail in the real world where missing data occur. It follows that the designer must plan to accommodate incomplete observations. In this article we consider in particular the design consequences of anticipated patterns of missing data as well as mechanisms, as a prelude to a separate discussion of methods of analysis of such datasets (see companion paper by Siddique,[4] page 793, in this issue).

An array of statistical methods have addressed some of the problems resulting from missing data, and high quality methods are now implemented in many software packages. For modeling the type of repeated measures data that often occur in psychiatric trials, the Laird-Ware mixed effects model[5] for conducting repeated measures analysis of variance, along with its direct and collateral descendents,[6-8] provides a set of powerful tools for analyzing longitudinal data. The uptake of these new methods into psychiatric research was accelerated by

publications[9] that described their benefits and strengths. One much-cited and appreciated benefit is the easy way that they accommodate arbitrary patterns of missing data, at least in the sense that the software runs and produces apparently sensible estimates and inferential statistics (means, standard errors, confidence intervals, and tests of null hypotheses). Because of these convenient statistical packages, inferences can be drawn without requiring the researcher to think hard about the analysis of datasets with partially missing data on subjects, including truncation by dropout as well as sporadic gaps as patients miss visits.

However, the fact that the software runs is not a guarantee that the output is sensible or valid. The validity of analytic results depends on the correspondence between the assumptions of the underlying model and the realities of the data, and to the sensitivity of those models to inevitable departures from the assumptions. Sometimes, a powerful analytic tool produces a “Type III error,” which is the correct answer to the wrong question.[10] In keeping with this issue's emphasis on important matters of concern to the psychiatric researcher at the time of planning and study design, this article deals with the conceptualization of missing data from a design point of view. Good design may not eliminate the problem of missing data, but as we will show, it can reduce it, so that the modern analytic machinery can be used to extract statistical meaning from study data. Conversely, we note that when insufficient attention is paid to missing data at the design stage, it may lead to inferential problems that are impossible to resolve in the statistical analysis phase.

The idea is summarized by the following “Method for Capturing Lions in New York City:” 1) by design, New York City does not harbor wild lions; and 2) the capture of tame lions is left as an exercise for the reader.

WHAT CONSTITUTES MISSING DATA?

A psychiatric researcher designing a randomized clinical trial of treatments must consider the nature and consequences of various mechanisms that might create gaps in the measurement record of some subjects. Our psychiatric researcher might begin by imagining what it would be like to have perfect power of observation. That is, suppose she is able to detect and measure whatever actually occurs, whenever it occurs (but not what does not occur). Are there still going to be gaps in the data? For example, suppose the study is designed around 8 weekly study visits, where the Hamilton Depression Rating Scale (HAM-D) Y is to be measured. If the patient dies of suicide after the fourth visit, there will be no Y's measured at visits 5 through 8. Is this a missing data problem?

Data must “be” to “be missing.” There is no post-mortem state of symptoms that can be captured in a score. The data are not “missing” because they do not correspond to real, definite states of nature. One might object that they have “potential” reality (what would the symptoms be if the patient had not died), but then they lack definiteness because we have not specified how the patient came not to die. Consider two of many possibilities: a) the doctor cures the patient of depression and thus averts the suicide, or b) the doctor commits the patient to involuntary hospitalization and the suicide is prevented by close observation. One would expect the symptom status to depend on how the suicide was averted.

The unavailability of data due to mortality may have only modest salience in psychiatric research due to the small risk of death in most studies. It is much more important in areas such as metastatic cancer research where mortality is a common outcome. However, even in psychiatric research there can be situations where non-being masquerades as missing data. For example, suppose a patient often has a panic attack while crossing a bridge, and Y measures the rate of panic attacks while crossing bridges in the past week. If the patient avoids bridges completely in a week, is Y missing, or zero, or undefined? The way the investigator conceptualizes the relationship between panic attacks and avoidant behavior will determine whether a missing data “method” is appropriate.

There are situations at the other extreme, where there is no question that data could exist, but is only unavailable due to the experimenter's action dictated by the study protocol. In the earlier example where a full diagnostic assessment was selected for only a subset of those screened negative, there is no question that their diagnoses could exist, and thus they are missing by design. Whenever the mechanism that causes this incomplete data is known, such as in this case where a random portion of screen negatives have missing data, then analyses of these data are likely to produce reliable inferences when the underlying model is well specified.[11]

We discuss two more subtle examples of missing by design. The first conceptualizes each subject as having two sets of possible outcomes, one set when that subject is assigned to treatment A, and the other when assigned to treatment B. Because each subject can only be assigned one intervention, the outcome on the other, nonassigned treatment is certainly missing. This perspective that each subject in a two-arm trial has two potential outcomes, only one of which can ever be observed, plays a fundamental part in the Neyman-Rubin-Holland causal modeling approach to experimental trials.[12-14] Again, if the mechanism for why certain subjects are observed on treatment A and others are observed on treatment B is well known, such as in a successful randomized trial, appropriate inferences about treatment effects can be obtained.

This concept of potential outcomes is not limited to the primary outcome of a trial but can also include side effects or serious adverse events resulting from a drug, or even patient behavior such as non-adherence. These postintervention measures can be conceptualized as two distinct outcomes, one under an active drug, for example, and one under a placebo. For every patient, we can observe only one of these measures; which one is observed is dependent on that patient's intervention assignment.

As another example of missing by design, consider varying the times that subjects are measured at follow-up. In principle, subjects could be assessed any day after beginning the study. Even though we typically use a specified number of windows in time to assess each subject's symptoms across time, in analyses we typically rely on a growth model of some form to represent this individual's symptom trajectory over the entire time interval. If we select a linear growth model, each person's trajectory can be represented by an intercept and slope, and these are considered to be random, or always unobserved, in our analyses. In principle, these random intercepts and slopes can also be treated as always-missing variables, that is, latent or unobserved variables. But do these intercepts and slopes have any ontological interpretation that can allow us to interpret them as missing data? We can at least conceive of a subject's profile of symptoms across all days since starting the study and drawing a best fitting line through all these points. The intercept and slope from this line could then be measured summaries of the beginning point and linear trend for that individual. Indeed, random effects and other always-missing or latent variables can indeed be conceived and handled as missing data in analyses.[15] There is something important that is given up, however, when there is complete missing data on a variable. The quality of the inferences rides on the adequacy of the underlying model assumptions involving these random effects or latent growth variables. It is also difficult to verify model fits because the important variables of inferential interest are unobserved.[1,16]

Other types of incomplete data arise in response to practical limitations in the study design. For example, for some designs, an assessment may only occur upon attendance at a treatment visit, a somewhat risky design choice compared to one where assessment is made regardless of a clinic visit. In a study with a psychotherapy arm, the patient who misses a treatment appointment would also miss that corresponding assessment. The only definite value for the missing data are the actual unobserved outcome in the patient under the circumstances that obtain. The outcome that would have occurred had the patient come to the treatment visit depends on the unspecified method for inducing the patient to adhere to treatment, and thus is not clear under this assessment protocol. A far safer design would be to separate assessment from treatment visit and conduct such assessments using a procedure that is blinded to intervention status. These options are not only different ways to measure the outcome, but also address distinct research goals.

Data that are censored can also be considered missing data. For example, if the outcome of interest is time to remission of depression, and the study ends before a subject remits, the remission time for that individual is censored, or known only to occur after the end of the last observation time. In this situation, the censoring mechanism, which is determined by the time at which the study ends, might not provide any additional information about the remission time. However, if a subject who improves decides to decline any assessments, then the time of censoring does provide information about remission time, and standard survival analysis can lead to improper inferences.

Truncated data are a special type of missing data that are distinct from, and often more difficult to model compared with censored or other types of missing data. Whereas for censored remission times, we know that there is a remission time for every subject in the study, even though some of these are censored before the study ends, for truncated data we do not even know of its existence. For example, in many trials there are limited records kept on those who decline to consent to be randomized. Without any data on these subjects, other than knowing perhaps the proportion of those approached who declines to consent, we cannot assume that the differences in response to the different treatments that we observe for consenters is the same as that for non-consenters. This type of truncation due to unknown self-selection factors would make it difficult to assess treatment impact at the population level.

Sometimes the researcher unintentionally truncates the sample due to the analytic method. Older statistical programs often deleted any cases where a subject had any missing data, so reports of impact using such crude analytical procedures were limited to inferences on those with complete information. Occasionally, a researcher will choose to report an analysis based only on those who complied with the treatment regimen, rather than the more traditional intent-to-treat analysis. Again, these inferences will only relate to a subset of the original population. Finally, truncation can occur naturally in group-based randomized trials, such as those involving whole school, classroom, or community interventions. In particular, students who enter after the intervention period begins or leave the school before the end of an intervention period may often be ignored in any follow-ups.[1]

The statistical theory underlying all modern missing data methods relies on the underlying reality or interpretability of the measures, (or subjects) that are treated as missing. The "missingness mechanism" specifies the probability of data being missing or a subject being truncated as a function of that individual's data (whether all observed or partly or wholly missing). When this mechanism is known, it is possible to account for this missing data in the model; when it is unknown and related to the data that are missing, inferences can be erroneous.[15] As an example, let us return to the bridge-avoiding issue, in which we have an experiment involving two interventions aimed at reducing panic attacks. In one condition the patients' rate of panic attacks decline over a few weeks to a low level (ie, one minor attack in 10 crossings). In the other condition patients' attacks are so severe that they avoid nearly all bridges, crossing just once or twice during the course of the study, only when sufficiently sedated that they do not experience panic attacks. Any statistical method that treats the response as missing for "no crossing" weeks will estimate 1/10 for the first condition and 0 for the second. Thus, the second condition will come out looking better on the statistical test of the null hypothesis. This erroneous result can be avoided by redefinition of the outcome variable — one choice is to create an ordinal variable with no missing data where the categories range from "avoid bridges entirely," "cross bridge with panic attack," and "cross bridge without panic attack." The right choice will depend on the intent of the investigator. The main message is that sweeping the problem into missing data has unintended consequences.

MINIMIZING MISSING DATA PROBLEMS IN INTENT-TO-TREAT ANALYSES

Analyses based on each patient's assignment to intervention condition, rather than by what they actually received, are called intent-to-treat (ITT) analyses, which have become the primary means used to compare intervention conditions in randomized trials. To be definite, suppose response variables are measures of the HAM-D at each of 8 weekly visits after randomization to either a standard antidepressant A or a novel compound B. Some patients randomized to receive B will stop taking it (go “off protocol,” fail to be “adherent,” or “dropout,” as this is variously described). Some of these “non-adherers” will substitute another off-protocol treatment (perhaps A), while others will go untreated. A similar situation holds for the patients who are randomized to receive A. In the ideal case, the researcher will continue to obtain the HAM-D scores that reflect the true state of symptoms that actually occur in all patients (no missing data). Given that non-adherence often occurs, what question does an intent-to-treat analysis answer? It is often described as measuring the relative effect of the two “policies:” try to give A versus try to give B. The actual adherence behavior of the patients is not of direct concern in addressing this question.

Now consider the case where there are missing longitudinal data and we still want to conduct ITT analysis. As long as missing data are unrelated to adherence status, there is no conceptual problem in obtaining an inference regarding this intent-to-treat question.[15] But if outcomes are missing whenever non-adherence occurs, then we can no longer separate the effects of these two assignment conditions from adherence on these two drugs.

In this context, it is important to distinguish between two distinct types of non-adherence:

1. Treatment non-adherence: patient not following the treatment protocol; and
2. Assessment non-adherence: patient not providing the assessment data.

Unfortunately this distinction is often over-looked, sometimes unintentionally, as some research teams are unaware of this distinction, and sometimes intentionally. In particular, many study design protocols have determined that assessment should end whenever there is non-adherence. This is a mistake, and it limits our ability to make sensible inferences from such studies.

The conduct of a trial can, and indeed has, greatly reduced this type of missing data. It has been observed that excellent follow-up rates can be obtained even in outpatient studies, if subjects are educated about the need to continue adhering to the measurement protocol (study visits), even if they choose to refuse to adhere to the treatment protocol. This produces less missing data overall, and therefore reduces our dependence on inferences on unverifiable assumptions about the missing data. The entire study team including the investigators, project directors, study coordinators as well as the assessment staff, should also be made aware of the need to separate the assessment from treatment protocol adherence. Such education is particularly necessary for staff members who have been trained on protocols calling for termination of subject measurement at protocol deviation.[17] Turning a non-adherence problem into a non-adherence and missing data problem is simply a mistake. Given that this investigator wants to follow the ITT principle, there is every reason to try to measure all outcomes in all subjects.

WHAT IS WRONG WITH LOCF TO ANSWER THE INTENT-TO-TREAT QUESTION?

In longitudinal trials, the most common pattern of missing data has been full dropout, where missing at one time point is followed by missing data at all following time points. Any approach to handling these dropouts can naturally only use data up to this last time point. One method has had a long history of being used to deal with dropouts, the LOCF imputation method. The deficiencies of this peculiar imputation method have been known for 2 decades,[17] but its use in clinical trial research publications and new drug applications has continued to be practiced widely. Only recently has the recognition of LOCF's problems spread widely in psychiatric research and regulatory agency thinking.

The original idea behind the widespread use of LOCF[17] was to “penalize” a treatment arm that was failing, by “freezing in” the typically poor observed response of patients in the experimental arm of early antipsychotic trials at the point that the experimental treatment was discontinued for intolerable side effects, gross lack of beneficial effect, or both. The argument was made that one could not make direct use of the post-discontinuation data, and that any biases were conservative (against the experimental drug). Thus, LOCF was conceived as a way to deal with the structural “dropout” created by protocol-driven cessation of experimental treatment, while ostensibly preserving the completeness of the longitudinal data. In this use, the main faults are the following:

1. The imputed post-discontinuation scores lack a definite unobserved underlying value. Because no one would imagine continuing the drug in the clinic in such a patient, there is no real value in knowing “what would have happened” if it had been continued. The LOCF does not impute what happened when the patient discontinued the experimental drug (and perhaps was switched to a standard treatment). LOCF is intended to substitute for the readily available post-discontinuation scores.

2. The assertion that LOCF provides conservative point estimates for treatment effects is questionable. The same argument that LOCF underestimates the expected treatment outcome for the experimental drug (even if valid) can also be made to argue that LOCF underestimates the expected treatment outcome for the control/placebo arm. Therefore it is not clear that the net result is biased toward underestimating treatment effects that contrast the two arms. This is especially problematic if the treatment non-adherence rate is higher in the control/placebo arm than in the experimental arm.

3. The completed LOCF dataset has statistical properties that cannot be handled using standard statistical models for longitudinal data. For example, correlations between observations, which must be taken into account, are heavily affected by discontinuations. In a trial with biweekly assessments, the correlation between sequential observations depends dramatically on the time the patient discontinues. If there is a discontinuation at 4 weeks, for example, all observations afterward in a LOCF dataset are perfectly correlated. If there is no discontinuation, the correlations are far smaller. Also, any variability in the true outcomes is ignored by LOCF, so the precision of the study is overestimated. Not only is this method unsound, but it is also not assured of producing conservative results as has been previously claimed.[18] The most serious effect of LOCF is that it lulls the designer into a false sense of security about the need to resolve complex issues of non-adherence, as well as the consequences of unrestrained missing data for precision and bias. This “moral hazard” has led investigators to underdesign studies in the mistaken belief that LOCF cures all.

What about using LOCF for interpolating sporadic missing data? The main point to take away is that whenever there is a well-defined objective for LOCF as an imputation method, there is a preferred analytic method that can do that job better. (See the article by Siddique et al, page 793, regarding detail on these preferred methods.) We would leave the reader with the “black box warning” that LOCF can be dangerous, can yield nonsensical or invalid results, and can always be safely replaced by a combination of redefined goals and sound analytic technique. Thus, LOCF has no place in the modern toolkit.

ALTERNATIVES TO ITT

The above comments relate to the investigator's choice to follow the ITT principle. But what if there is nonadherence and we want to go beyond the goals of the ITT analysis? The most common reason to want something other than the ITT "policy comparison" is the desire to come to grips with non-adherence to the study protocol. Most investigators, given the choice, would naturally prefer to compare treatments under ideal conditions of control and adherence. Some investigators go so far as to discount the value of the postadherence outcomes, thinking that once a subject departs from the treatment protocol there is no reason to value or even use the subsequent data. This has led in the past to a habit of truncating follow-up at non-adherence. As we have discussed above, this creates a missing data problem of a particularly vicious kind (see article by Ten Have, page 772).

There has been a great deal of activity in the area of alternatives to ITT that are based on randomization (in the way they make inferences) and do not require heavy assumptions.[19] Such methods include "principal stratification," and the "instrumental variables" or "complier average causal effects" models, and they are beginning to see use in placebo controlled trials. We briefly present the ideas underlying these methods here. All of these methods generally presume that the target population is heterogeneous; for example, a proportion of the population would experience side effects serious enough to cause nonadherence, were they to be put on a particular drug. Would this subgroup experience potential harm on the drug while the remaining adherers generally benefit? One way of assessing impact is to present findings separately for groups based on their adherence status, but unfortunately, information on adherence is only partly available. In the active drug arm of a trial, we are able to identify those individuals who adhere or do not adhere to this drug. We cannot tell how they would adhere under the placebo condition. Similarly, in the placebo arm, we cannot determine what they would actually do on the active treatment arm; we can only assess whether they would adhere to the placebo condition. The principal stratification and related methods provide the means for assessing impact on symptom and other outcomes, conditional on a cross-classification of the population on adherence in drug and placebo conditions. Such a cross-classification results in four classes: a group that would adhere to both active drug and placebo, a group that would adhere to neither, and two groups that would adhere to one but not the other. Although the full cross-classified status of each person can never be determined directly, inferences about treatment benefit or harm can be made for each of these classes once appropriate restrictions in the model are made. [20]

It should be emphasized that all the modern ideas that go beyond ITT make use of postadherence observations and cannot generally be used unless there is at least some postadherence observation. There is also a strong theoretical connection between missing data methods and "beyond ITT" methods. But the connection between the desire to go beyond ITT and the missing data created by truncation of postadherence entirely artificial, caused by a misunderstanding that creates a poor design choice, resulting in a self-inflicted wound.

The methods that go beyond ITT (with as complete follow-up of what actually happens as possible) represent an advance over plain ITT, but they are not a cure-all for the lack of experimental control. It would take us beyond the scope of this review to deal adequately with the rich theoretical and practical developments of the past decade. However, in view of the particular importance of the issue of non-adherence, and its historical connection to phenomenon of missing data "by design," we mention a set of methods and approaches that go by the loose appellation "designing for ITT." This allows treatment to adapt to individual level preferences as well as responses under previous treatments.[19]

DESIGNING FOR ITT

The basic idea is that the ITT principle provides solid randomization-based inference about policies, but sometimes the comparison of those policies does not satisfy the investigator's needs. Instead of trying to eke out inference related to “what if” questions from the observed data, we investigate ways to reframe the treatment comparison a priori (in the design) so that the resulting design provides a better fit to the investigator's research goals when the resulting data are analyzed by ITT. For example, suppose the investigator seeks to evaluate the promise of a new antidepressant against a standard antidepressant, recognizing that this standard may be satisfactory for a subset of depressed patients. If one were to use a standard design, one might employ a three-way randomization among placebo, a standard antidepressant, and the new drug, with some weeks of follow-up. We can anticipate that this design could lead to a number of nonadherents. First, the placebo patients might “drop in” at high rates to one of the antidepressants because the intervention is not effective. Secondly, the standard treatment might cause unpleasant side effects, also causing patients to switch or abandon treatment altogether. If the new treatment is well-tolerated, and no worse in effect than the standard, it might achieve high levels of adherence. The ITT analyses would compare what actually happens in all three assigned groups (ignoring adherence, assuming perfect observation). The comparison with placebo is particularly unsatisfactory, because we are really comparing the standard (or the innovation) with a mixture of placebo and “rescue” medications. It is hard to know what to do with the results, especially if placebo “looks good” on ITT but there are high rates of drop-in.

An alternative design might provide a clearer answer to another policy relevant question, particularly if the new treatment is intended as a second-line alternative to a standard treatment, when there is only a partial or no response to the standard treatment. In that case, the relevant comparison is to begin the trial with everyone treated by a single standard antidepressant and then randomize to the new antidepressant or to a second standard when there is nonresponse under this first drug. One might also treat all patients with the standard, and then randomize patients who do not have a complete response to innovation-based therapy, switching, or augmenting, depending on whether the response is partial or not at all, or to standard-based therapy again switching or augmenting from the current armamentarium. Here the placebo plays little or no role. Then the ITT analysis is precisely what is needed. Thus, the idea is to convert trials whose ITT analysis

is unsatisfactory into new designs that are suited to ITT (see the companion paper by Ten Have, see page 772, for other alternatives to ITT).

CONCLUSIONS

An aim of this article has been to tame the problem of missing data, by removing the “wild” parts. These wild parts are:

1. The confusion between structurally missing observations (postmortem quality of life, and other solecisms) and probabilistically missing data, where there is an unobserved but definite reality that could be observed in principle.

2. The truncated dataset that results from a misguided attempt to work around the ITT principle when the investigator wants more than a comparison of policies.

We recommend that investigators deal with structural matters by careful definition of outcomes to avoid structural missing data, whenever possible. We also suggest that when adherence is likely to be far from perfect and ITT does not seem relevant, investigators consider a redesign to a study for which ITT is satisfactory. We also suggest that investigators explore one of the modern methods for going beyond ITT. We urge investigators not to compound an adherence problem with a self-inflicted missing data problem. In the end, what counts is faithful observation of well-controlled data; that is what separates experiment from anecdote.

We also recommend four general principles for reducing problems related to the remaining type of “tame” missing data.

Minimize the rate of missing data

All the inferential problems with missing data, and especially those that flow from failures of assumptions in models, leverage off the missing data rate. Thus, whatever can be done to bring that rate down will reduce the need to use unverifiable assumptions about the missing data.

Avoid missing data that depend on some unobserved value

This second rule, if it is followed, will yield the most benign form of missing data, whose consequences are merely a loss of precision due to reduced information, without bias in (for example) treatment comparisons. For example, the two-stage follow-up design that uses a first stage screen for everyone followed by a more expensive diagnosis, was based on an observed variable — the screen — and random selection. Such data missing by design are easy to correct for with modern statistical techniques. On the other hand, if a clinic visit is required to obtain assessment data, such a visit may be too onerous for those who are deeply depressed, or unnecessary for those who are feeling well. Either way, the chance that a subject's assessment is missing may depend on his or her symptom level.

Seek information about the reasons for missing data and the observable patient characteristics that predict it

This rule helps to verify that missing data are not related to unobserved variables. Additional information allows one to adjust for the observable part of measurement bias. For example, suppose patients who have bad side effects tend to drop the study treatment, and then tend to do poorly on the primary symptom scale as a result. The ITT principle will count that outcome against the treatment, of course, and it is important to observe it. However, we may imagine that despite efforts to the contrary, patients who have decided to drop treatment for side effects may be less likely than their adherent colleagues to come in for the study visit that would reveal their increased symptomatology. In that scenario, the (observed) side effects profile creates a connection between missing data and the unobserved outcome. If one can (appropriately) include the side effects measures in the outcome modeling, this part of the bias can be reduced.

Seek proxies for the missing data

This last rule allows one to recover somewhat from failures of the previous two rules. Efforts might include telephone contacts to supplement (or replace) a standard visit, or the use of informants, utilization databases, retrospective “make up” sessions to recall the recent past when visits were missed, and so on. Most analytic methods will produce a “predicted” or “model-based” estimate of the expected outcome at a particular visit. These can be compared with the results of proxies. If the discrepancies of predicted and proxy vary systematically according to the observation status, the assumptions of the model may be questioned.[21]

Once the missing data have been tamed by design, and by following the rules described above, inference in the presence of the modest amount of garden variety missing data can be managed by one of the many efficient and well-studied modern methods described in a companion article (see Siddique et al, page 793). As always, the dose makes the poison, so our warnings apply with force that depends on the amount of such missing information. At 5% or even 10% of tame missing data in a trial, there is not likely to be any important difference among methods. As the proportion of missing data rise, and particularly if the missing data mechanism has a substantial wild component, the choice of methods becomes more delicate. To avoid depending on statistical witchcraft, the investigator will do well to invest in design choices that minimize the wildness and the scope of missing data — this is no more than good technique.

We have concentrated on experiments (clinical trials), but some of the same issues come up in observational studies, and the investigator has less control over them. One major new problem concerns missing confounders (predictors of treatment and outcome that are unobserved). Rubin[22] points out that the design of observational studies can benefit from considering some of the important issues that motivate familiar trial design features, such as pre-specification of the analytic model. The absence of experimental control should not imply the freedom to ignore the prospect of missing data. To the contrary, all other problems of observational studies are sharpened by missing data, so the designer must remain no less alert to the issues described above.

REFERENCES

1. Brown CH, Wang W, Kellam SG, et al. the Prevention Science and Methodology Group. Methods for testing theory and evaluating impact in randomized field trials: intent-to-treat analyses for integrating the perspectives of person, place, and time. Drug Alcohol Depend 2008;95(Suppl 1S):S74–S104. [PubMed: 18215473]
2. Newman SC, Shrout PE, Bland RC. The efficiency of two-phase designs in prevalence surveys of mental disorders. Psychol Med 1990;20(1):183–93. [PubMed: 2242109]Erratum in: *Psychol Med.* 1990;20(3):following 745
3. Shrout PE, Newman SC. Design of two-phase prevalence surveys of rare disorders. Biometrics 1989;45 (2):549–555. [PubMed: 2765638]
4. Siddique J, Brown CH, Hedeker D, et al. Missing data in longitudinal trials – Part B Analytic Issues. Psychiatric Annals 2008;38(12):793–801. [PubMed: 19668352]
5. Laird N, Ware J. Random effects models for longitudinal data. Biometrics 1982;38(4):963–974. [PubMed: 7168798]
6. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. Biometrics 1994;50(4):933–944. [PubMed: 7787006]
7. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal probit and logistic regression analysis. Comput Methods Programs Biomed 1996;49(2):157–176. [PubMed: 8735023]
8. Hedeker D. A mixed-effects multinomial logistic regression model. Stat Med 2003;22(9):1433–1446. [PubMed: 12704607]
9. Gibbons RD, Hedeker D, Elkin I, et al. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. Application to the NIMH treatment of Depression Collaborative Research Program dataset. Arch Gen Psychiatry 1993;50(9):739–750. [PubMed: 8357299]
10. Kimball AW. Errors of the third kind in statistical consulting. Journal of the American Statistical Association 1957;52(278):133–142.
11. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. Wiley; New York: 2002.
12. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 1974;66:688–701.
13. Rubin DB. Bayesian inference for causal effects: the role of randomization. Annals of Statistics 1978;6:34–58.

14. Holland PW. Statistics and causal inference. Journal of the American Statistical Association 1986;81:945–960.
15. Hedeker, D.; Gibbons, RD. Longitudinal Data Analysis. Wiley; New York: 2006.
16. Carlin JB, Wolfe R, Brown C, Gelman A. A case study on the choice, interpretation, and checking of multilevel models for longitudinal, binary outcomes. Biostatistics 2001;2(4):397–416. [PubMed: 12933632]
17. Lavori PW. Clinical trials in psychiatry: should protocol deviation censor patient data? Neuropsychopharmacology 1992;6(1):39–48. [PubMed: 1571068]with discussion and comment 49-63
18. Molenberghs G, Thijs H, Jansen I, et al. Analyzing incomplete longitudinal clinical trial data. Biostatistics 2004;5(3):445–464. [PubMed: 15208205]
19. Goetghebeur T, Loeys E. Beyond intention to treat. Epidemiologic Reviews 2002;24:85–90. [PubMed: 12119861]
20. Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics 2002;58(1):21–29. [PubMed: 11890317]
21. Laird NM. Missing data in longitudinal studies. Stat Med 1988;7(12):305–315. [PubMed: 3353609]
22. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2006;26(1):20–36. [PubMed: 17072897]